知识产权信纸设计-01

专利技术交底书模板

*了方便事务所与您的沟通，以使您的专利申请尽早完成，烦请填写下表。*

|  |  |  |  |
| --- | --- | --- | --- |
| **专利申请技术交底书** | | | |
| 发明人  基本  资料 | 第一  发明人  (对本发明做出最大贡献的人) | 姓 名 | 王吉驰 |
| 身份证号（中国大陆居民提供） | 342423199004267376 |
| 其它  发明人(如有) | 姓 名 |  |
| 申请人  基本  资料 | 第一申请人（即谁拥有这个专利） | 名 称 | 飞算云创 |
| 地 址 |  |
| 其它  申请人(如有) | 名 称 |  |
| 地 址 |  |
|  | | | |
| 如申请中国大陆专利，以上信息需用中文 | | | |

**专利申请技术交底书（软件类）**

**（红色标示部分必须提供）**

**技术问题联系人：** 王吉驰 **联系人电话：** 15850000988 **联系人e-mail: 1046304858@qq.com**

1. **发明名称**

IV值和相关性分析方法在数据科学领域中对类别与连续型目标特征的数据泄露检验

**二、与本发明最接近的现有技术及其缺点**

本发明在数据科学领域, 尤其构建风控模型时可以更为简单自动化的识别出泄露数据的变量，运用多种统计方法，组合，并确定相关阈值，可以利用其高效率，高精准的识别具有数据泄露风险的变量。传统的检测数据泄露风险的方法主要有以下几个缺点，首先没有统一检验标准，其次更需要依赖人为主观筛选，最后整个过程较为耗时，而且效率低下。具体操作是先对所有变量与目标变量通过相关性矩阵计算找到相关性强的特征，再进一步通过数据探索，如数据对比或者分组箱体图。最后通过人工区分观察，找到有数据泄露风险的嫌疑特征。这种方式需要人工介入观察，如果特征特别多，例如超过10000以上，这样逐个排查便非常耗时繁琐，还容易出错，处理起来非常麻烦。在本发明中，可以实现全自动化筛选操作，数据泄露识别方法更为多样化，并且在效果上更为精准。

**三、本发明所要解决的技术问题**

1、数据获取、清洗和特征处理，这里主要是针对于获取到的原始数据，首先缺失值，字符型数据的处理，并且涉及到缺失率的计算，和数据类型的识别。主要包括非结构化数据的整理和转换为二维DataFrame，以及文本类数据的标签化处理等。

2、目标类特征的处理，考虑到目标类特征可能是连续值或者离散值，并且对于离散值可能是多分类也可能是二分类，这里对于离散目标特征统一做二分类处理。

3.1、对于目标变量为离散型的处理方式：

3.1.1、离散型目标变量所对应的特征变量，自动化处理这些特征得到最优分箱，识别自变量的对应类型，分别对连续变量按照等值或者等距进行分组，然后设定阈值，如特定得卡方值，达到阈值后停止分箱，离散列直接按照特定得卡方阈值进行分箱合并，最终得到每一列得最优分箱。

3.1.2、离散型目标变量所对应的特征分箱结果调整并计算最终的IV值。首先根据每个特征不同分组计算对应得WOE值，然后根据分组WOE的趋势是否单调来对箱体进行合并处理。并最终根据调整后的分箱计算每个对应特征的IV值。

3.1.3、IV阈值的设定，对于目标变量为离散型的特征计算IV值之后由大到小进行排序，设定相应的阈值，并对大于阈值的特征变量进行自动化数据审核。

3.2、对于目标变量为连续变量的处理方式：计算特征变量与目标变量的相关性系数，设定对应的阈值，并对大于阈值的特征变量进行自动化数据审核。

4、集成树模型模型结果验证与特征重要性审查，对于存在数据泄露的特征，如果参与了模型训练，那么一定会存在模型的训练和验证结果都非常好。并且树模型可直接得到入模特征的重要性排序，因此可通过之前步骤检验结果，根据是否让存在数据泄露风险的特征入模训练，比较前后训练结果和对应特征重要性排序结果，进一步验证数据泄露风险特征。

**四、本发明的技术方案详描**

**1、数据获取、清洗、特征处理**

本次研究数据获取来源主要由银行信用卡客户数据，包括用户个人基本信息，用户申请表数据，以及还款计划表数据，第三方渠道数据，以及公务系统的征信数据等构成。根据客户ID将获取到的数据进行并表处理，这里我们默认拿到的还款计划表已经经过客户的ID号分组合并之后的数据。之后对于并表后的数据进行相应的数据清洗，主要包含对各个特征变量的类型识别属于离散变量还是连续变量或者属于时间类型，对缺失率过多的列进行识别和剔除，对日期列进行转换，单一值列数据进行相应的剔除，并且对错乱的字符串离散列列进行剔除。对文本类数据如文字描述等数据按照正面、负面、中性进行标签化处理，最后得出初步成型的中间结果数据。同时依据中间结果数据做初步的描述性统计分析，对于离散型特征数据不同分组计算样本个数，对于连续性特征分别计算均值方差和各分位数等。此处的数据清洗，和初步特征的处理可以筛选出相当一部分特征，为后面相应的相关性指标和IV值计算减少计算次数。

1. **目标特征的处理**

目标特征的处理包含分类目标变量缺失处理、多分类目标变量的处理、以及连续目标变量的处理等。

缺失目标变量的处理，首先根据用户的ID号进行去重处理，并且将目标变量列所对应的缺失值直接按照行剔除对应的样本。

多分类目标变量的处理，依据缺失目标变量的处理结果，对目标变量进行分组统计样本个数，并按照统计样本的个数进行降序排列，排序第一的作为一类，剩余的所有组作为另一类处理，最终将多分类转换为2分类。

连续目标变量的处理，直接依据缺失目标变量的处理结果，将缺失值行剔除。

**3.1、离散型目标变量的数据泄露检验方式**

**3.1.1、自动化处理特征变量**

如果预测目标变量是离散型，那么对于其对应的特征变量首先根据每个特征的数据类型按照分类变量和连续变量进行划分，对于日期类特征变量，如果了解实际业务含义首先做该特征做对应的特征衍生转换为连续型变量然后在进行处理。

连续型特征变量，一般首先按照等值或者等距的分割方式进行默认初步分组，分组之后，再利用卡方分割的方式，设置最小卡方值，一般默认为最小卡方值默认为3.84，通过循环将各组进行自动化合并并达到最终的最优分箱。如果连续型特征内含有缺失值、奇异值、或者离群值，对于缺失值首先按照均值或者中位数等填充，然后再进行等值，或者等距方式进行分割，考虑到前面已经对于缺失率超过50%以上的特征，已经做过剔除处理，如果对于有实际业务含义的缺失值，此时也可以将缺失值作为单独一箱进行处理。对于奇异值，或者离群值等特殊值，如果同样有实际业务含义，在划分分组时候也可以单独作为一组处理。

离散型特征变量，对于缺失、特殊值等按照类似于连续型变量的处理方式，如果有特殊的含义则单独分为一箱。对于其余的组别，可依据卡方分箱的方式，按照最小卡方值得到最优分箱。

**3.1.2、对于分箱结果进行自动化调整，并且计算调整之后的IV值**

特征变量在自动化分箱之后，对于每个特征的不同分组都可以依据目标变量计算WOE的值，设置横坐标为分组的下标，纵坐标为每个分组WOE的值，于是对于每一个特征变量，都可以得到不同分组的WOE条状图，据此可以得出WOE的趋势，程序上主要通过遍历每个特征的所有分组，按照分组的顺序比较每个分组前后值的大小并计数，最终根据计数值大小可得该特征WOE的整体趋势。如果该特征对应的WOE趋势完全单调，那么不需要对该特征继续做分箱处理。如果对应的趋势不单调，则根据整体的趋势，将趋势发生改变的箱体所对应相邻的箱进行合并处理处理并重新计算WOE的值，直到最终WOE趋势单调，则分箱调整截止。根据WOE单调趋势调整后的分箱，再根据IV值的计算公式计算每一个特征字段的IV值。最终将所有的特征计算出的IV值进行排序，并保存为对应的字典格式。

**3.1.3、IV阈值的设定与自动化数据审核**

IV阈值的设定，我们已经对于目标变量为分类变量的特征计算IV值之后由大到小进行排序，并且依据IV值的计算公式，一般情况下，IV值可以正常计算，但是如果在分组中出现了响应数为0或者非响应数为0的情况。此时计算出来分组的woe就为负无穷，得到的IV值为正无穷。这个时候需要将这样特殊的分组单独做成一个规则作为补充。因为IV衡量的是某一个变量的信息量，其值的大小决定了自变量对于目标变量的影响程度。大部分计算的IV值分布在0.02和0.5之间，本发明确定将IV值小于0.02的特征剔除，不再进行参与后续流程，将IV值大于0.5的特征记录并考虑剔除，因为这部分的特征包含目标的信息量过多，可能存在数据泄露的风险，需要进一步进行处理。

IV值大于0.5之上的特征进行数据审核处理，这里的数据审核主要是分别查看目标变量在这特征变量的分布情况。如果在这些特征有较为明显的区分，则判断该特征存在数据泄露可能.此处程序化验证区分情况，主要通过统计特征分组内好坏样本的个数，如果存在多个分组内好或者坏样本的个数为0的情况，则判定为具有明显区分。

**3.2、连续性目标变量的数据泄露检验方式**

目标变量类型是连续型。考虑到每个特征变量和目标之间都有相关性关系，所以如果特征变量是离散型，那么此处使用spearman 相关性系数方法计算，如果特征变量是连续型，则对应的使用person 相关系数方法。最终我们可以得到每个特征变量与目标变量之间的相关性关系。一般相关系数的值在 0.8-1.0 之间认为极强相关，系数在0.6-0.8 强相关，表示强相关，0.4-0.6 中等程度相关，0.2-0.4弱相关，0.0-0.2极弱相关或无相关。所以这里将相关系数阈值设定为0.8。考虑到当相关系数大于0.8，这个时候特征变量蕴含大部分目标变量的信息，因此该特征变量可能存在数据泄露风险，同时记录该特征。

**4、集成树模型模型结果验证与特征重要性审查**

对于目标变量不同类型的特征，以上步骤分别通过IV值筛选与数据审核检验的方式以及相关性阈值筛选的方式初步找到了存在数据泄露可能性的特征。紧接着需要进一步通过集成树模型的验证，本发明调用的是Lightgbm方法，主要考虑到该集成树方法兼容性强，可直接处理缺失值，处理离散变量。我们将处理后的离散字段重新编码。最终将所得结果按照训练集和测试集按照7比3进行分割，将所有的特征入模调用Lightgbm算法，我们可以得到模型训练后的结果。对于结果首先我们观察是否训练和验证集对应的AUC都超过0.95以上，如果是，则再查看对应排名前十的特征重要性系数指标，如果同样非常靠前，在归一化处理之后非常接近1，并且在之前IV大于0.5或者相关性大于0.8以上的记录特征内，那么则判断该特征具有数据泄露风险。然后将具有潜在数据泄露风险的特征剔除，再次调用Lightgbm算法，并查看对应的训练集和验证集的AUC，是否出现一定程度的下降，如若没有，可能仍然存在数据泄露的字段，需要下调对应阈值，直到最终模型训练和验证结果的AUC有较为明显的下降，如下降幅度超出0.05以上，则最终识别出所有潜在的泄露字段。

**五、上述第四项中的某些步骤是否有其他的实现方式**

考虑到算法的执行效率，和数据集的大小以及算法的兼容性和健壮性这里主要是指可以对缺失值等进行处理，本发明调用的是LightGBM算法，针对于模型算法的调用一般树模型的方法都可以采用，因为几乎所有的树模型算法都可以得到特征重要性质指标。如果对于性能要求不是很高也可以调用XGBoost模型进行相应的处理。

本发明在自动化最优分箱的过程中，考虑到分箱的效率，和分箱方法的兼容性，采用的是最小卡方值分箱。初步分箱后进行相应的合并操作，当达到设置最小的卡方值时，则自动化分箱停止，这里也可以通过其他统计指标，如初始等频或者等距分箱之后，然后通过递归的方式依据IV值是否有提升，对箱体进行合并操作。如果合并之后IV值并无明显的提升，则在分箱合并停止，并记录最优分箱，但是这种自动化最优分箱方式需要通过一次循环并且里面嵌套递归的方式完成，当数据量和数据特征变多的时候，会带来较大的性能问题。

**六、描述本发明的有益效果。**

本发明高效且精准的实现了自动化数据泄露检验的功能，更智能化的对数据泄露特征识别。通过自动化数据清洗，以及相应的统计指标计算如缺失率，单一值个数等，将不符合要求的特征进行剔除操作从而优化数据质量，提升验证效率。对于目标变量的处理多样，自动化检测目标变量类型，并对缺失目标值进行剔除处理和对多分类目标值二分类处理。使得后续步骤可兼容IV值计算，提升整个检验流程的健壮性。通过设定最小卡方值自动化处理分箱，并根据WOE趋势调整分箱结果可使得特征可解释性更强。将IV值较大的特征记录为潜在风险泄露特征，并进一步通过数据审核和相关树模型算法验证结果和对应的特征重要性，大大提升了数据泄露检验的准确性。考虑到连续型目标变量，又提供了初始相关性检验的方式，也大大提升了数据泄露检验的多样性和范围，使得整个流程可处理更多错综复杂的数据泄露问题，从而得到较为精准的结果。

**七、需要保护的关键技术点总结**

1、文本类数据处理，首先分别构建相应的正面负面和中性的词库，然后通过jieba分词，和TFIDF的方式提取本文关键词，最后通过判断分出来的关键词属于哪一种词库范围内将该数据重新转换为正面、负面和中性。

2、目标特征的处理，这里主要针对不同类型的特征做相同或者异同的处理。

相同处理方式

步骤1：首先目标变量缺失值进行剔除操作。

步骤2：按照用户的ID列进行去重处理。

不同处理方式

对于多分类变量，首先对各个类别进行分组统计样本个数，并且按照计数大小由大到小进行降序排列，计数最多的作为一类，剩余的组作为另一类从而将多分类转换为二分类进行处理。

连续型目标变量则只将缺失值对应的行进行剔除操作。

3.1、目标变量为离散型，其数据泄露的检验方式

3.1.1 特征变量的自动化最优分箱，这里将特征变量按照离散和连续分别进行处理。

对于离散型特征变量

按照卡方分箱方式进行处理

步骤1：首先计算每个离散变量对应离散值的样本占比，以及每个分组内坏样本的占比

步骤2：其次根据每个分组对应样本占比的大小进行重新编码

步骤3：通过设置最小卡方值，将编码后的各个组别依照坏样本占比相邻的组不断循环进行合并处理，并且计算合并后对应分组的卡方值，如果达到阈值则停止合并，最终得到最优卡方分组。

对于连续性特征变量

步骤1：按照等值或者等频的方式进行初始分箱。并对分箱后的组别进行相应的编码处理

步骤2：对于初始分组后的特征同样按照类似于离散变量最小卡方值分箱的方式进行合并处理，并最终得到最优分箱结果

至此所有特征变量都已经分箱完毕，设定最小卡方值，整体只需要经过一次循环，这样可以提升整体分箱速度，而且通过卡方分箱的方式，可降低后续分箱调整的复杂程度。

3.1.2、自动化调整分箱，并计算特征变量的IV值

自动化最优分箱完整之后，需要进一步依据WOE的趋势进行相应的分箱调整

步骤1：根据最优分箱的结果计算对应特征每个分组的WOE值

步骤2：计算每个特征变量对应分组的WOE趋势是否单调

步骤3：对于奇异点，WOE非单调的组和相邻的组进行合并处理

步骤4：重新计算每个特征新分组的WOE的值并验证其是否单调

步骤5：对于最终调整分箱后的特征计算对应的IV值并保存，供后续筛选使用

对于WOE趋势的判断，遍历每个特征的所有分组，按照分组的顺序比较每个分组前后值的大小并计数，例如[0.1,0.2,0.21,0.1,0.4]，遍历该WOE值列表，0.2大于0.1，记录数加1，0.21大于0.2记录数再次加1，继续遍历0.1小于0.21，记录数减去1，0.4大于0.1，记录数继续加1，最终记录数3大于0，则该特征整体WOE趋势单调递增，但是第三次遍历的时候，出现偏差，这个时候将第三和第四组合并，然后再重新计算各组的WOE值，使得所有分组WOE趋势单调。最终根据调整后的分组，计算对应的IV值，方便后续筛选数据泄露特征。

3.1.3、IV阈值的设定以及自动化数据审核

对于IV解释预测能力，相关研究表明IV值小于0.02则表示该特征无预测能力，IV值在0.02和0.1之间表示该特征预测能力较弱，在0.1和0.3之前表示预测能力中等，超出0.3则表示预测能力很强。实际数据预测过程中发现IV值超出0.5的特征，对于的预测能力极强，所以本发明所确定的IV筛选数据泄露风险的阈值为0.5。IV值计算完成之后，可根据IV值的大小进行相应的排序，将IV值小于0.02的特征剔除，不再参与后续流程，记录IV值大于0.5的特征，这些特征可能存在数据泄露风险，进一步对于对这些潜在风险泄露字段做数据审核，验证目标变量在该特征上面的分布情况，如果有明显的区分，则该特征存在潜在数据泄露风险。

3.2、目标变量为连续型，其数据泄露的检验方式

步骤1：通过person相关系数计算方法计算连续型特征变量与目标变量的相关系数

步骤2：通过spearman相关系数计算方法计算离散型特征变量与目标变量的相关系数

步骤3：根据设定得相关系数阈值0.8，将超出该阈值得特征记录，做进一步得探索

步骤4：根据设定得相关系数阈值0.2，将低于该阈值得特征剔除，不参与后续流程

4、集成树模型结果验证与特征重要性审查

步骤1：将之前步骤处理后的数据按照训练集和验证集进行相应得分割

步骤2：调用Lightgbm算法得到模型结果AUC的值和特征的重要性系数

步骤3：根据模型的结果，如果训练和验证集的AUC都超过0.95以上，则大概率存在潜在数据泄露风险，然后再观察特征重要性排序，看排序靠前的特征是否在之前步骤IV筛选或者相关性判断记录的潜在数据泄露风险特征中

步骤4：将之前步骤潜在的数据泄露特征剔除，然后再次调用Lightgbm算法并查看对应的模型结果是否出现一定的下降通常AUC下降幅度在0.1以内，且结果稳定即未出现过拟合情况。

至此综合以上流程，即可自动化得出潜在数据泄露特征，并且检验效果精确，稳定。后续相应的建模人员或者业务人员便可根据检验的结果进行特征审核，最终判定是否存在数据泄露特征。

**八、请列举本发明已知和潜在的技术/产品应用领域及其应用方式。**

信用风险评估，信用卡评分，精准营销，数据泄露检验，模型监控。

**注意：**

1. **英文缩写要注明其英文全文和中文译文。**
2. **全文对同一事物的叫法应统一，避免出现一种东西多种叫法。**
3. **在上述功能模块中，若其功能是现有技术的模块中不能实现的，则提供的内容不能只有原理，也不能只做功能介绍，需要提供该功能模块的内部结构。**知识产权信纸设计-01